ตอน
3

สำหรับตอนที่ 3 นี้จะกล่าวถึง เทคนิคทั่วไป เกี่ยวกับตัวอักษรและรหัสภาษาไทย ที่มีผลต่อการพัฒนาทางซอฟท์แวร์ต่อไป

เทคนิคพัฒนาการ ภาษาไทย บนเครื่อง ไมโครคอมพิวเตอร์

ยืน ภู่วรวรรณ

ตัวอักษรไทยที่ใช้งานปัจจุบัน

ตัวอักษรไทยที่ใช้กันในปัจจุบัน จะต้องได้รับการกำหนดรหัสขึ้นแทน รหัสที่ใช้แทนนี้หากพิจารณาดูบนตาราง ISO646-1983 ซึ่งเป็นตาราง ASCII แล้ว เราสามารถขยายเพิ่มต่อได้อีกเท่าตัว(128 อักษร) แต่อย่างไรก็ตาม เครื่องคอมพิวเตอร์บางเครื่องรับรหัสเพียง 7 บิทเท่านั้น การขยายเพิ่มต่อจึงเสมือนการใช้หลักการ shift ซึ่งเมื่อเป็นเช่นนี้ รหัสภาษาไทยที่น่าจะใช้ได้ จึงต้องหลบรหัสควบคุมบางส่วนออกไป โดยเริ่มจากรหัส A0 เป็นต้นไป

ตารางของรหัส ISO646-1983 เป็นดังตารางที่ 1

				b7	0	0	0	0	1	1	1	1
				b6	0	0	1	1	0	0	1	1
				b5	0	1	0	1	0	1	0	1
b4	b3	b2	b1		0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	@	P		p
0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	2	STX	DC2	"	2	B	R	b	r
0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
0	1	0	0	4	EOT	DC4	¤	4	D	T	d	t
0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
1	0	0	0	8	BS	CAN	(	8	H	X	h	x
1	0	0	1	9	HT	EM	)	9	I	Y	i	y
1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
1	0	1	1	11	VT	ESC	+	;	K	[	k	{
1	1	0	0	12	FF	IS4	,	<	L	\	l	|
1	1	0	1	13	CR	IS3	−	=	M	]	m	}
1	1	1	0	14	SO	IS2	.	>	N	^	n	‾
1	1	1	1	15	SI	IS1	/	?	O	_	o	DEL

ตารางที่ 1 รหัส ISO0646 ซึ่งกำหนดรหัสไว้ครบแล้วใน 7 บิทแรก

สำหรับตัวอักษรไทยทั้งหมด หากจะพิจารณาดูจะประกอบด้วยตัวอักษรที่สำคัญดังนี้

พยัญชนะ ได้แก่

ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ฤ ล ฦ ว ศ ษ ส ห ฬ อ ฮ (พยัญชนะมีทั้งหมด 44 ตัว กับ ฤ และ ฦ สำหรับพยัญชนะ ฃ และ ฅ นั้น ตามพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2525 กล่าวไว้ว่า "เดี๋ยวนี้เลิกใช้แล้ว" แต่เพื่อเป็นการอนุรักษ์มรดกไทย ก็อาจเก็บไว้ในลำดับเดิมและมีการกำหนดรหัสให้ แต่การป้อนข้อมูลผ่านทางคีย์บอร์ดอาจทำไม่ได้ จึงต้องป้อนข้อมูลวิธีอื่น เช่น การกำหนดรหัสโดยตรง)

สระ ได้แก่ ็ ะ ั า ำ ิ ี ึ ื เ แ
ุ ู ํ (นฤคหิต) และ ฺ พินทุ รวม

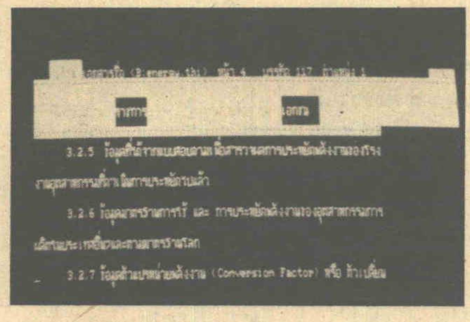
วรรณยุกต์และการันต์ ได้แก่ ่ ้ ๊
รวม 5 ตัว

เครื่องหมายพิเศษที่ใช้งานทั่วไป

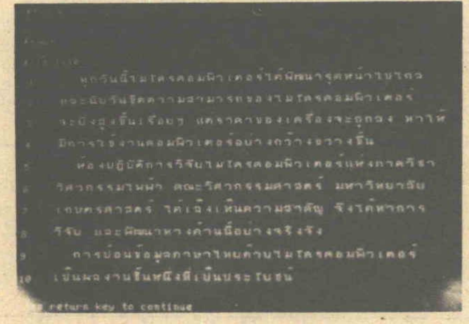
เครื่องหมายพิเศษเฉพาะ ◉ ฟองมั
โคมูตร ๛ ยามักการ ฿ เครื่อ
เงินบาท รวมเครื่องหมายพิเศษ

เลขไทย ได้แก่ ๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗
รวม 10 ตัว

เป็นตัวอักษรที่ใช้ทั้งหมด 46 + 18
7 + 10 รวม 86 ตัวอักษร

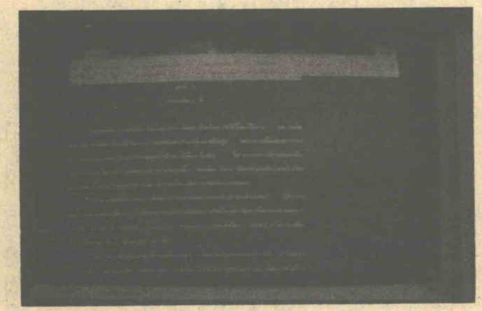


ภาษาไทยในระบบ 8 บรรทัด สามารถร่นเป็น 12 บรรทัดได้ นภาพเป็น Thaieasy ซึ่งจะแสดงเอกสาร 5 บรรทัด แต่เมื่อกด H จะลบเมนูออกแสดงได้ 7 บรรทัด (ในเวอร์ชัน 4.1) การจัดการ อภาพถ้าเราเขียนโปรแกรมทำงานเองโดยไม่ผ่าน OS จะทำให้ สามารถทำงานได้เร็วมาก (การ scroll)

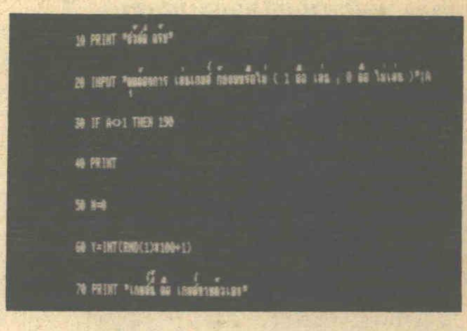


```
    ทุกวันนี้ไมโครคอมพิวเตอร์ได้พัฒนารุดหน้าไปมาก
และมีขีดความสามารถของไมโครคอมพิวเตอร์
จะยิ่งสูงขึ้นเรื่อยๆ แต่ราคาของเครื่องจะถูกลง ทำให้
มีการใช้งานคอมพิวเตอร์อย่างกว้างขวางขึ้น
    ห้องปฏิบัติการวิจัยไมโครคอมพิวเตอร์แห่งภาควิชา
วิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัย
เกษตรศาสตร์ ได้เล็งเห็นความสำคัญ จึงได้ทำการ
วิจัย และพัฒนาทางด้านนี้อย่างจริงจัง
    การป้อนข้อมูลภาษาไทยด้านไมโครคอมพิวเตอร์
เป็นผลงานชิ้นหนึ่งที่เป็นประโยชน์
 return key to continue
```

ภาษาไทยด้วยระบบกราฟิก เราสามารถใช้ซอฟท์แวร์ร่วมกับ CRTC จัดการกราฟิกในรูปแบบภาษาไทยได้ ในภาพเป็นการใช้ ราฟิกบนเครื่อง Textronic 4051 ซึ่งสามารถจัดจำนวนบรรทัด ามที่ต้องการได้ (ขึ้นกับความละเอียดของภาพ)

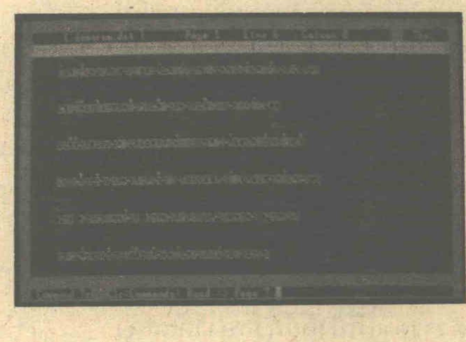


ขวัญใจนักพิมพ์ดีด เวอร์ชัน 4.5 เป็นการใช้วิธีการจัดระบบ วยตัวโปรแกรมเอง ทำให้ไม่มีปัญหายุ่งยากเหมือนการเข้าไปแก้ไข อฟท์แวร์จากต่างประเทศในระดับภาษาเครื่อง (ในภาพเป็นการใช้ อร์ดโมโนโครมหรือโมโนโครมกราฟิก โดยความเร็วของการจัด กรีนบนจอสูงมาก)

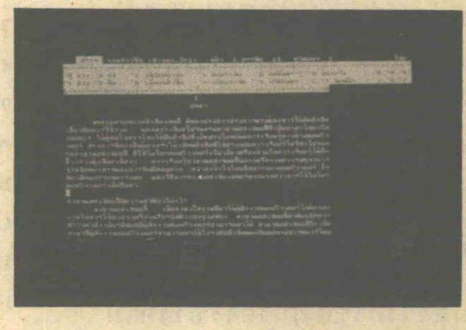


```
10 PRINT "[illegible]"
20 INPUT "[illegible] ( 1 [illegible] , 0 [illegible] )";A
30 IF A<>1 THEN 190
40 PRINT
50 N=0
60 Y=INT(RND(1)*100+1)
70 PRINT "[illegible]"
```

เมื่อใช้ TSX เชื่อมกับ BIOS แล้ว เราสามารถใช้ภาษาเบสิกของไมโครซอฟท์ให้ทำงานร่วมได้อย่างธรรมชาติ ในภาพเป็นการใช้ TBASIC ร่วมกับ MBASIC แต่ถ้าใช้ Thai.com ใน IBM และต้องรัน MSBASIC ซึ่งเป็นเบสิกแบบ line editor (ไม่ใช่แบบ full screen เหมือน BASICA)



ในบางเครื่องสามารถทำ font load screen ได้ เช่นในเครื่อง Victor หรือ Apricot ทำให้เราสามารถสร้างตัวอักษรอะไรก็ได้ หรือจะกำหนดคีย์บอร์ดเป็นอะไรก็ได้ ในภาพเป็นการนำเอา Thai easy เวอร์ชันแรก ๆ รันเป็นเวิร์ดโปรเซสซิ่งสมัยพ่อขุนรามคำแหง



ภาพภาษาไทย 25 บรรทัด ที่แสดงนี้เป็นการประยุกต์โครงสร้างซอฟท์แวร์อีกวิธีหนึ่ง จัดการโดยไม่ต้องแก้ไขฮาร์ดแวร์แม้แต่น้อย

ภาพสั่นไหวลดลง แต่ต้องแก้ไขและเพิ่มเติมฮาร์ดแวร์ ซึ่งเมื่อเทียบกับการใช้วิธีอื่นที่ดัดแปลงฮาร์ดแวร์น้อยกว่า วิธีนี้จึงต้องลงทุนทางฮาร์ดแวร์สูง เพราะต้องทำบอร์ดอะแดปเตอร์เพิ่มเติมเข้ามา

ตัวอักษรไทยกับคีย์บอร์ด

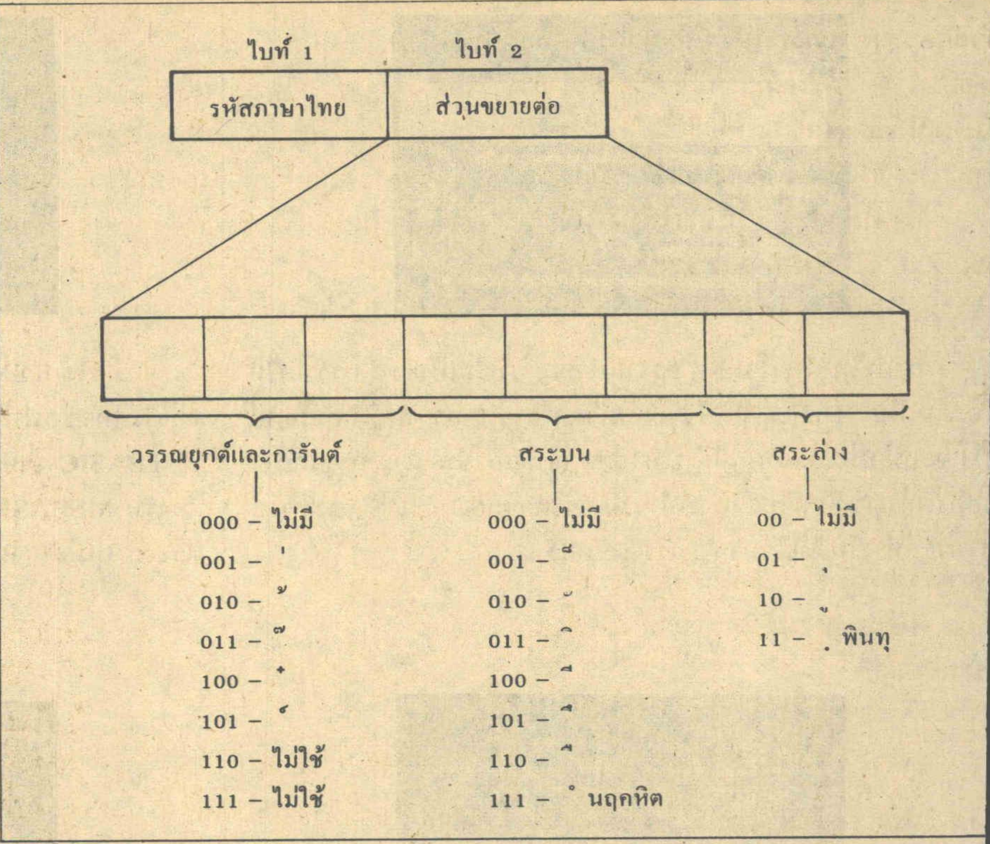
คีย์บอร์ดของคอมพิวเตอร์ก็ใช้การจัดวางเหมือนกับพิมพ์ดีดไทยทั่วไป การจัดวางคีย์บอร์ดของเครื่องพิมพ์ดีดไทย มีสองรูปแบบคือ รูปแบบแบบดั้งเดิมมา และแบบที่นักวิจัยชื่อ สฤษดิ์ ปัตตะโชติ ได้วิจัยและเสนอแนะ โดยรัฐบาลก็ยอมรับและเคยประกาศให้หน่วยงานราชการใช้ แต่ความคุ้นเคยของผู้ใช้ก็ยากที่จะแก้ไขและเปลี่ยนแปลงได้ จนในปัจจุบันการจัดวางคีย์แบบเดิมมีผู้ใช้มากกว่า 90% ของจำนวนผู้ใช้ และเคยมีผู้สำรวจเมื่อเร็วๆ นี้ว่า มีผู้ใช้คีย์บอร์ดแบบปัตตะโชติเพียงประมาณ 8% ของทั้งหมด และมีแนวโน้มไม่เพิ่มขึ้น เพราะแม้แต่แบบเรียนสอนพิมพ์ดีดของกระทรวงศึกษาธิการก็กำหนดแป้นพิมพ์แบบเดิม

การจัดวางตำแหน่งคีย์ทั้งสองได้แสดงดังรูปที่ 3 และ 4

สำหรับการจัดแป้นพิมพ์ทางด้านไมโครคอมพิวเตอร์หรือเทอร์มินัลนั้น มักจะอยู่ภายใต้การควบคุมของซอฟท์แวร์ ดังนั้นจึงแก้ไขและเปลี่ยนแปลงให้เป็นอะไรก็ได้ โดยเฉพาะอย่างยิ่งผู้เขียนซอฟท์แวร์ จะเขียนซอฟท์แวร์ด้วยภาษาแอสเซมบลีร่วมกับอินเตอร์รัพท์รันคีย์บอร์ด แล้วแก้ไขสร้างตารางการปรับคีย์ เพื่อสร้างคีย์บอร์ดเป็นอะไรก็ได้

ข้อพิจารณาทางภาษาศาสตร์กับการพัฒนา

โครงสร้างการพัฒนาการทางด้านซอฟท์แวร์ จำต้องอาศัยหลักวิชาการและอาศัยประสบการณ์ประกอบร่วมกัน จึงจะทำให้ระบบซอฟท์แวร์นั้นมีประสิทธิภาพยิ่ง โครงสร้างของภาษาไทยเป็นสิ่งที่เป็นธรรมชาติ จนยากที่จะหากฎเกณฑ์ตายตัวได้



รูปที่ 1 การจัดรหัส 2 ไบท์

		1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0																
1																
2	SPAู่CE	กู่	ขู่	คู่	ฆู่	งู่	จู่	ฉู่	ชู่	ซู่	ฌู่	ญู่	ฎู่	ฏู่	ฐู่	ฑู่
3	ฒู่	ณู่	ดู่	ตู่	ถู่	ทู่	ธู่	นู่	บู่	ปู่	ผู่	ฝู่	พู่	ฟู่	ภู่	มู่
4	ยู่	รู่	ฤู่	ลู่	วู่	ศู่	ษู่	สู่	หู่	ฬู่	อู่	ฮู่				
5	SPAู้CE	กู้	ขู้	คู้	ฆู้	งู้	จู้	ฉู้	ชู้	ซู้	ฌู้	ญู้	ฎู้	ฏู้	ฐู้	ฑู้
6	ฒู้	ณู้	ดู้	ตู้	ถู้	ทู้	ธู้	นู้	บู้	ปู้	ผู้	ฝู้	พู้	ฟู้	ภู้	มู้
7	ยู้	รู้	ฤู้	ลู้	วู้	ศู้	ษู้	สู้	หู้	ฬู้	อู้	ฮู้				
8																
9																
A	SPAู๊CE	กู๊	ขู๊	คู๊	ฆู๊	งู๊	จู๊	ฉู๊	ชู๊	ซู๊	ฌู๊	ญู๊	ฎู๊	ฏู๊	ฐู๊	ฑู๊
B	ฒู๊	ณู๊	ดู๊	ตู๊	ถู๊	ทู๊	ธู๊	นู๊	บู๊	ปู๊	ผู๊	ฝู๊	พู๊	ฟู๊	ภู๊	มู๊
C	ยู๊	รู๊	ฤู๊	ลู๊	วู๊	ศู๊	ษู๊	สู๊	หู๊	ฬู๊	อู๊	ฮู๊				
D	SPAู๋CE	กู๋	ขู๋	คู๋	ฆู๋	งู๋	จู๋	ฉู๋	ชู๋	ซู๋	ฌู๋	ญู๋	ฎู๋	ฏู๋	ฐู๋	ฑู๋
E	ฒู๋	ณู๋	ดู๋	ตู๋	ถู๋	ทู๋	ธู๋	นู๋	บู๋	ปู๋	ผู๋	ฝู๋	พู๋	ฟู๋	ภู๋	มู๋
F	ยู๋	รู๋	ฤู๋	ลู๋	วู๋	ศู๋	ษู๋	สู๋	หู๋	ฬู๋	อู๋	ฮู๋				

รูปที่ 2 ตัวอย่างตารางการกำหนดรหัส 2 ไบท์ส่วนหนึ่ง จะเห็นว่าในหนึ่งแบลงค์ (256 ตัวอักษร) ได้แบ่งเป็น 4 page การเลือกแบลงค์และ page จะใช้รหัสไบท์ 2 เป็นตัวเลือก ในภาพเป็นการสร้างการ escape มายัง page ต่างๆ ใน blank ตามรูปแบบของไบท์ที่สอง ที่กำหนดมา(ถอดรหัสโดยไบท์ที่ 2)

ตัวอักษรดังกล่าวนี้ จึงกล่าวได้ว่าเป็นตัวอักษรสำหรับแลกเปลี่ยนข้อมูลภาษาไทย แต่อย่างไรก็ตาม ตัวอักษรหลายตัวอาจมีผู้แย้ง ว่า ในเมื่อเลิกใช้แล้วจะเก็บไว้ทำไม คำตอบ สำหรับกรณีนี้เห็นจะได้แก่ มีไว้ไม่เสียหลาย

ปัญหามีอยู่ว่า ในการใช้งานไมโคร อมพิวเตอร์ทั่วไป จำเป็นจะต้องใช้รหัส สดงผล และเทคโนโลยีปัจจุบันยังไม่เอื้อ านวยต่อการแสดงผลแบบกราฟิกที่ใช้รหัส ดี่ยวได้ การใช้รหัสผสมจึงเป็นความจำเป็น หัสผสมนี้จะทำให้การแสดงผลทั้งบนจอ าพและเครื่องพิมพ์ดูเป็นธรรมชาติขึ้น และ นเวลาการทำงานในกรณีของเครื่องพิมพ์ จะต้องพิมพ์ถึง 4 ระดับมาเป็น 3 ระดับ

ตัวอักษรผสมที่พิจารณานำมาใช้ควร ระกอบด้วย

ํ่ ํ้ ํ๊ ํ๋ ั่ ั้ ั๊ ั๋ ิ่ ิ้ ิ๊ ิ๋ ิ์ ี่

ี้ ี๊ ี๋ ึ่ ึ้ ึ๊ ึ๋ ื่ ื้ ื๊ ื๋

รหัสผสมที่จำเป็นนี้ มี 25 ตัว แต่ใน รแสดงผลเฉพาะบางอย่าง อาจนำตัวอักขระ งตัวกระจายออก เพื่อให้แสดงผลได้สวย มขึ้น แต่ก็ต้องเปลืองรหัสมากขึ้น เช่น

พ รวมกับ ˌ เพื่อให้ได้ ฟ ที่หางยาวขึ้น

เ รวมกับ เพื่อให้ได้ ไ มีส่วนสูงมากขึ้น

(ที่สำคัญคือ ส่วนล่างของ ไ คือ เ จะใช้ร่วมกับสระ เ ไม่ได้ เพราะส่วนจุดการแสดงผลจะขาดไป ทำให้ไม่สวยงาม จึงต้องแยกรหัส ไ ออกเป็น 2 ตัว)

การกำหนดรหัสลงในตารางจึงควร หลักการที่สำคัญคือ รหัสที่กำหนดเมื่อใช้ เครื่องที่รับ 7 บิท แล้วใช้หลักการ shift มีผลกระทบน้อยที่สุด นอกจากนี้รหัสที่ หนดนี้ จะกระทบต่อรหัสควบคุมเฉพาะ งอย่างของเครื่องคอมพิวเตอร์น้อยที่สุด มกระทบกับโปรแกรมจัดระบบงาน กระทบ โปรแกรมประยุกต์อื่น ๆ

ระผสมมีความสำคัญเพียงไร

สระผสมในที่นี้หมายถึง การนำรูป ะสองตัวมารวมกันเพื่อกำหนดเป็นรหัสหนึ่งตัว เช่น ิ่ รหัสผสมมักใช้สำหรับการแสดงผลด้วยหลักการภาษาไทย 3 บรรทัด ซึ่งต้องการรหัสเพิ่มขึ้นอย่างต่ำ 25 ตัว(หรือมากกว่า) ความสำคัญของรหัสผสมจึงเน้นเฉพาะการใช้งานในเรื่องการแสดงผล เช่น แสดงผลบนจอภาพ แสดงผลบนเครื่องพิมพ์ รูปแบบของรหัสผสมนี้ จะช่วยทำให้การจัดการภาคแสดงดีขึ้นและสะดวกขึ้น เช่นการพิมพ์ตัวอักษรของเครื่องพิมพ์ดอทเมทริกซ์ สระผสมจึงไม่มีความจำเป็นโดยตรง กับงานแลกเปลี่ยนข้อมูลและข่าวสาร ไมโคร–คอมพิวเตอร์เครื่องหนึ่งอาจแสดงผลด้วยอักษรสระผสมได้ แต่เก็บข้อมูลในหน่วยความจำแบบแยกกัน เช่น ิ ่ เป็น 2 รหัส

การขยายรหัสสำหรับอักษรภาษาไทย

ในการแสดงภาษาไทยนั้น มีตัวอักษรที่จำกัดอยู่เพียง 86 ตัวดังที่กล่าวแล้ว แต่เราอาจต้องการรหัสที่เป็นการผสมตัวอักษรอีก 25 ตัว ซึ่งก็ยังพอบรรจุลงในตารางขนาด 128 ช่องได้

ภาษาไทยเป็นโครงสร้างที่จัดตัวอักษรเป็น 4 ระดับ ซึ่งการประมวลผลต่าง ๆ จะต้องมีซอฟท์แวร์จัดการอยู่ด้วยเสมอ ทำให้ซอฟท์แวร์จากต่างประเทศ มักมีปัญหาในการทำงานกับข้อมูลภาษาไทยในระบบจัดระดับโดยตรง ทั้งนี้เพราะตำแหน่งทางฟิสิคัลที่มองเห็น กับตำแหน่งทางลอจิคอลที่อักษรอยู่ในหน่วยความจำมักไม่ตรงกัน การกำหนดตำแหน่ง การกำหนดรูปแบบการประมวลผลจึงมีปัญหามากเช่น ก ุ ้ ง และ ก ้ ุ ง ในทางลอจิคอลแล้วต่างกัน แต่ทางฟิสิคัลแล้วเหมือนกัน เป็นต้น

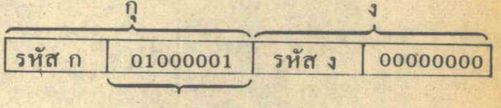
ผู้เขียนได้ลองขยายรหัส 2 ไบท์ เพื่อกำหนดฟิสิคัลกับลอจิคอลให้ตรงกัน แล้วจัดการกับระบบภาษาไทย 25 บรรทัด โดยยึดหลักการง่าย ๆ ดังนี้

ไบท์ 1	ไบท์ 2
รหัสภาษาไทย	ส่วนขยายต่อ

โครงสร้างรหัสสองไบท์

ลักษณะการกำหนดรหัสนี้ จึงเหมือนกับการแสดงผลในไอบีเอ็มพีซีอยู่แล้วคือ ไบท์แรกเป็นรหัส ASCII ไบท์ที่สองคือ แอตตริบิ้ว โดยใช้ไบท์ที่สองนี้ขยายรหัสออกอีก จึงเท่ากับว่าเราสามารถกำหนดรหัสภาษาไทยได้ถึง 65536 ตัว โดยมีหลักการกำหนดดังรูปที่ 1

ดังนั้น ถ้าหากจะแสดงคำว่า กุ้ง จะใช้ 4 ไบท์ ดังนี้



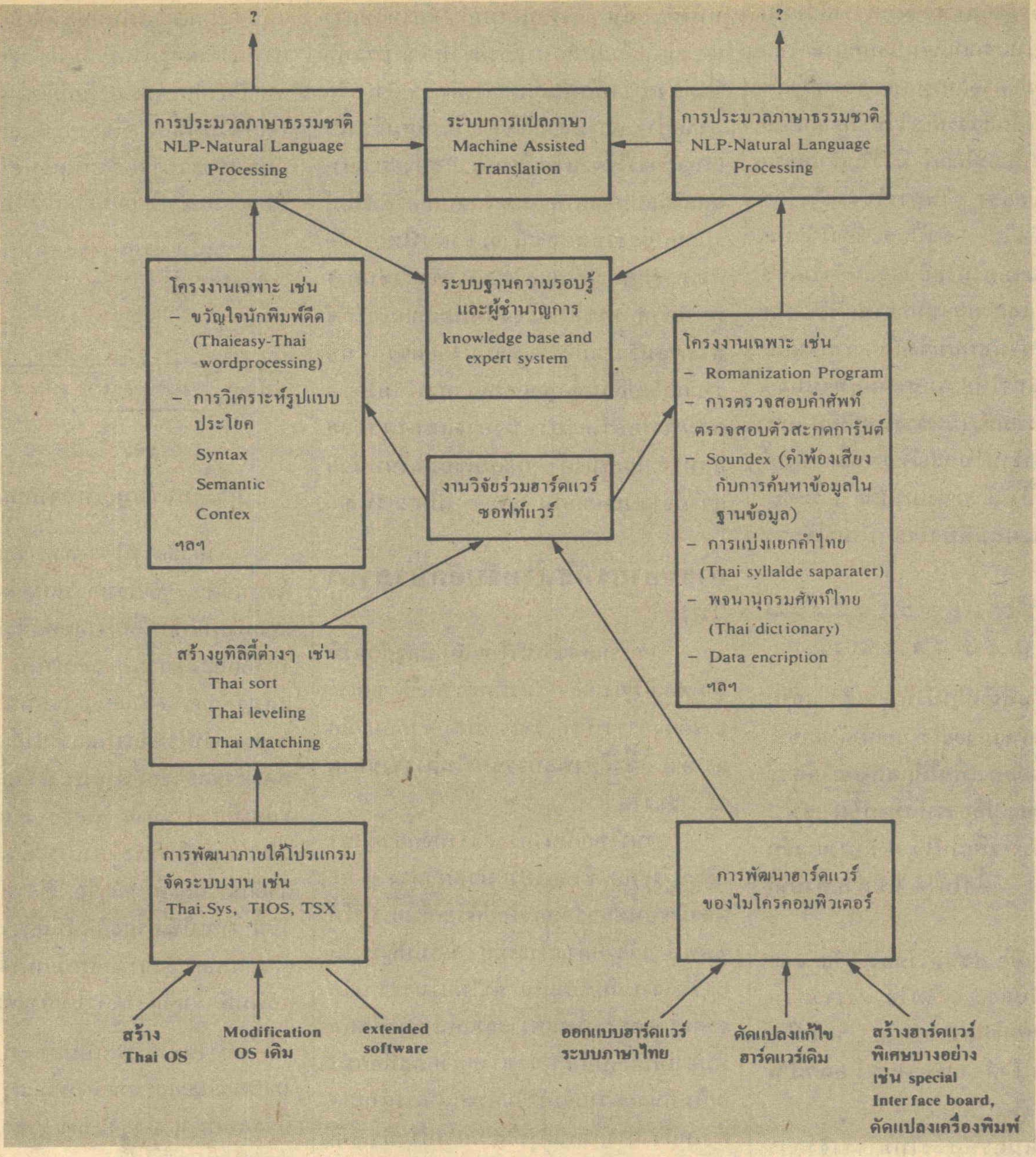
รหัส ้ และ ุ

การแทนข้อมูลด้วยรหัสสองไบท์

ด้วยวิธีการนี้ เราสามารถสร้างรหัสได้ครบทุกตัว สำหรับการแสดงผลภาษาไทย และหากใช้เทคนิคการลดคำ โดยใช้คำที่เป็นไปได้และมีความหมายเท่านั้น เช่น กุ้ คือ แสดงทั้งสระล่างและสระบนพร้อมกันไม่มีเราจะลดจำนวนตัวรหัสลงไปได้มาก จากการทดลองและวิจัยที่ผ่านมา ผู้เขียนพบว่าจะใช้รหัสเพียงประมาณ 2,800–3,200 ตัวก็พอเพียง ซึ่งอยู่ในวิสัยของการใช้ ROM กำเนิดตัวอักษร และที่สำคัญคือ วิธีการแสดงผลด้วยวิธีนี้ จะทำให้เครื่องพิมพ์ที่มีความละเอียดสูง (ในอนาคต) สามารถพิมพ์ภาษาไทยได้เร็วขึ้น คือพิมพ์ครั้งเดียวได้ครบทุกระดับได้โดยตรง

การขยายรหัสเป็นสองไบท์นี้ ผู้เขียนได้เสนอแนะและทดลองทำดูแล้ว เป็นวิธีหนึ่งที่จะลดปัญหาข้อยุ่งยากของซอฟท์แวร์จากต่างประเทศลงไปได้ เพราะซอฟท์แวร์ต่างประเทศจะเป็นรหัส 1 ไบท์กับแอตตริบิ้ว เราจึงใช้เสมือนเป็นแอตตริบิ้วในโหมดแสดงผลภาษาไทยได้ และใช้เทคนิคการสวิทช์ทางฮาร์ดแวร์ ช่วยให้แอตตริบิ้วเดิมทำงานได้ ในรูปที่ 2 เป็นตัวอย่างของการกำหนดรหัส 2 ไบท์ส่วนหนึ่ง

อนึ่ง: การพัฒนาด้วยรหัสสองไบท์นี้ฝ่ายพัฒนาของบริษัทถาวรคอมพิวเตอร์ได้พัฒนา ปรากฏว่าได้ผลดีในเรื่องความเร็วของการแสดงผลสูงมาก ความคมชัดและอาการ


?
?
การประมวลภาษาธรรมชาติ NLP-Natural Language Processing
ระบบการแปลภาษา Machine Assisted Translation
การประมวลภาษาธรรมชาติ NLP-Natural Language Processing
โครงงานเฉพาะ เช่น
– ขวัญใจนักพิมพ์ดีด (Thaieasy-Thai wordprocessing)
– การวิเคราะห์รูปแบบประโยค
Syntax
Semantic
Contex
ฯลฯ
ระบบฐานความรอบรู้และผู้ชำนาญการ knowledge base and expert system
โครงงานเฉพาะ เช่น
– Romanization Program
– การตรวจสอบคำศัพท์ ตรวจสอบตัวสะกดการันต์
– Soundex (คำพ้องเสียงกับการค้นหาข้อมูลในฐานข้อมูล)
– การแบ่งแยกคำไทย (Thai syllalde saparater)
– พจนานุกรมศัพท์ไทย (Thai dictionary)
– Data encription
ฯลฯ
งานวิจัยร่วมฮาร์ดแวร์ซอฟท์แวร์
สร้างยูทิลิตี้ต่างๆ เช่น
Thai sort
Thai leveling
Thai Matching
การพัฒนาภายใต้โปรแกรมจัดระบบงาน เช่น Thai.Sys, TIOS, TSX
การพัฒนาฮาร์ดแวร์ของไมโครคอมพิวเตอร์
สร้าง Thai OS
Modification OS เดิม
extended software
ออกแบบฮาร์ดแวร์ระบบภาษาไทย
ดัดแปลงแก้ไขฮาร์ดแวร์เดิม
สร้างฮาร์ดแวร์พิเศษบางอย่าง เช่น special Inter face board, ดัดแปลงเครื่องพิมพ์

รูปที่ 5 แสดงงานวิจัยที่เกี่ยวข้องกับการประมวลผลข้อมูลภาษาไทย

จำนวนคำที่ป้อน	จำนวนคำที่ได้(เก็บใน DICTIONARY)
10000	1826
20000	2557
30000	3094
40000	3532
50000	4002
60000	4297
70000	4608
80000	4889
90000	5028
100000	5165
110000	5216
120000	5286
133568	5337

ตารางที่ 2 แสดงจำนวนคำที่ใช้ในชีวิตประจำวัน

ั้นข้อมูลสนับสนุนบางอย่างจึงมีความ
ัญอย่างยิ่ง โดยเฉพาะงานพัฒนาการภาษา
บนไมโครคอมพิวเตอร์ มิได้อยู่ที่เพียงทำ
าร์ดแวร์แสดงผล ซอฟท์แวร์รับรู้ข้อมูลได้
ประมวลผลภาษาไทยมีแขนงแยกออกไป
ีกมาก เพราะภาษาไทยเป็นภาษาที่มี
ลักษณ์เฉพาะตัว มีความหมายและโครง
งประโยคที่แตกต่างกับภาษาอื่น มีไวยา-
ณ์และการใช้งานที่เหมือนภาษาธรรมชาติ
ปที่หากฎเกณฑ์แน่นอนไม่ได้

การพัฒนาการใช้ภาษาไทยด้วยไม-
คอมพิวเตอร์ จึงต้องกระทำต่อเนื่องเพื่อ
กิดประโยชน์ต่อการใช้งานในประเทศ
เราอย่างกว้างขวาง ห้องปฏิบัติการวิจัย
ครฯ ได้จัดโมเดลของการวิจัยออกเป็นรูป
ดังรูปที่ 5

งานวิจัยหลายอย่างจำเป็นต้องอาศัย
าการพื้นฐาน หรือข้อมูลที่จำเป็น ผู้เขียน
ะคณะร่วมทีมได้ทำการวิจัยหาข้อมูลพื้น
ข้อมูลเหล่านี้จะใช้สำหรับการวิเคราะห์
ูลแนะนำมาใช้ต่อไปได้ การวิเคราะห์หา
ิข้อมูลดังกล่าวมีจุดมุ่งหมายดังนี้

1. หาสถิติจำนวนคำศัพท์ที่ใช้ในชีวิต
ะจำวัน

2. แจกแจงความถี่ของคำไทย

3. แจกแจงความถี่ของการใช้ตัวอักษร

4. สถิติเกี่ยวกับรูปแบบคำไทย และ
จัดวางตัวอักษรตัวต้นและตัวท้าย เพื่อ
ข้อมูลแบ่งแยกคำไทย

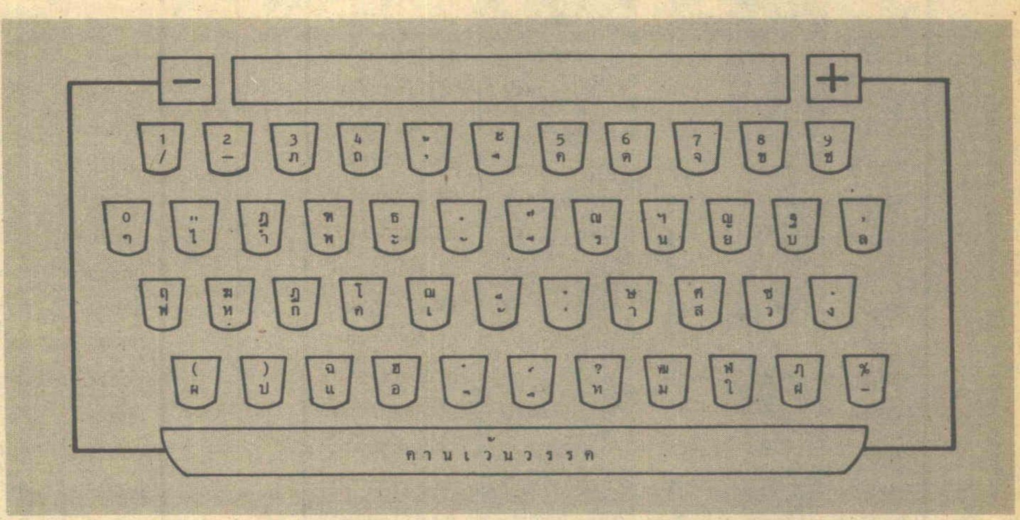
5. สถิติของคู่ตัวอักษรและตัวอักษร
บกล้ำ

6. สถิติขนาดคำไทย

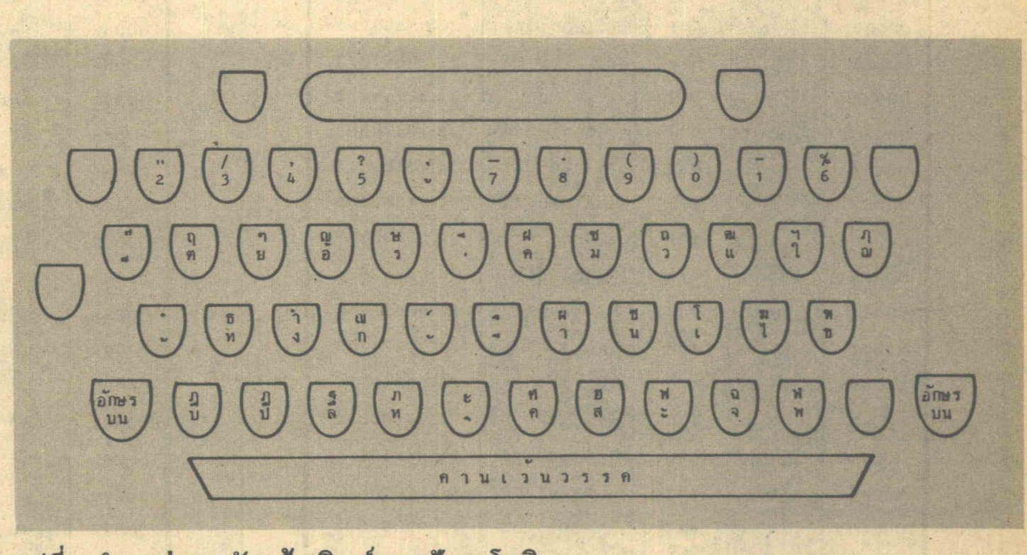
ข้อมูลที่ได้เหล่านี้ ดำเนินการมาโดย
ทำการสุ่มตัวหนังสือและเอกสารต่างๆ
ได้แก่ หนังสือพิมพ์ วารสาร นิตยสาร
เก็ตบุ๊ค รายงาน จดหมายราชการ หนังสือ
ทั่วไป โดยยกเว้นหนังสือประเภทวรรณ-
และตำราวิชาการที่แปลมาจากต่างประเทศ
ณะของการสุ่ม พยายามใช้หลักการทาง
ิ ตรวจสอบความเชื่อถือของข้อมูลตลอด
า

เอกสารเหล่านี้ จะได้รับการป้อนเข้ายังคอมพิวเตอร์ในลักษณะเป็นคำ โดยพยายามยึดหลักหน่วยคำหรือการแบ่งแยกด้วยมือเป็นคำ แล้วป้อนข้อมูลเข้าไปชุดละ 10,000 คำ ทำการจัดเรียงข้อมูลของคำ และ mearge ข้อมูลที่ซ้ำกัน จะเหลือเฉพาะคำและความถี่ โดยกระทำการเพิ่มข้อมูลเข้าทีละ 10,000 คำ เข้าไปยังฐานพจนานุกรมที่เก็บ และทำให้จำนวนคำเพิ่มขึ้นเรื่อยๆ จนกระทั่งจำนวนคำไม่เพิ่มมากนัก โดยตัวเลขคำศัพท์ที่เพิ่มขึ้นทีละหมื่นคำเป็นไปดังตารางที่ 2

หลังจากนั้นนำคำที่เก็บได้และความถี่มาจัดเรียงเพื่อดูสถิติของคำว่าคำไทยคำใดใช้มากที่สุด มีสถิติการใช้อย่างไร จากการวิจัยพบว่า คำไทยคำว่า “ที่” มีสถิติการใช้สูงสุด คือใช้ในความถี่ถึง 2.6% หรือกล่าวได้ว่าในคำไทย 100 คำ มีคำว่าที่ 2.6 คำ การจัดเรียงคำไทยโดยเรียงตามลำดับ 200 คำแรก ดังตารางที่ 3 และจากสถิติสะสมนี้ จะพบว่าจำนวนคำเพียง 200 คำนี้ มีสถิติเปอร์เซ็นต์สะสมถึง 67% หรือกล่าวได้ว่าใน 200 คำนี้ใช้ในชีวิตประจำวันถึง 64% แล้ว และจากการวิจัยพบว่า ในคำไทยเพียง 800 คำ จะใช้ในชีวิตประจำวัน 80% หรือถ้าต้องการให้เอทีเอ็ม (ATM) ของธนาคารพูดกับมนุษย์รู้เรื่องได้ 80% ในชีวิตประจำวัน เอทีเอ็มจะต้องเก็บคำศัพท์ไว้ 800 คำ ดังกล่าว



รูปที่ 3 ตำแหน่งการจัดแป้นพิมพ์ตามแบบเรียนสอนพิมพ์ดีดไทย พช 121



รูปที่ 4 ตำแหน่งการจัดแป้นพิมพ์แบบปัตตะโชติ

ตารางที่ 3 การแจกแจงความถี่ของคำที่ใช้ในชีวิตประจำวัน

ลำดับที่	คำไทย	ความถี่	เปอร์เซ็นต์	เปอร์เซ็นต์สะสม
00001	ที่	3474	2.60092837	2.60092237
00002	การ	2726	2.04090800	4.64183037
00003	เป็น	2073	1.55201844	6.19384881
00004	ได้	1936	1.44944896	7.64329777
00005	จะ	1874	1.40303066	9.04632848
00006	ใน	1766	1.32217297	10.36850140
00007	มี	1727	1.29297436	11.68147576
00008	ไม่	1568	1.17393387	12.83540963
00009	ก็	1552	1.16195495	13.99736458
00010	ของ	1541	1.15371945	15.15108403
00011	ให้	1486	1.11254192	16.26362595
00012	วา	1471	1.10131169	17.36493764
00013	ไป	1442	1.07959990	18.44453754
00014	และ	1423	1.06537494	19.50991248
00015	มา	1244	0.93136080	20.44127328
00016	ความ	1159	0.86772280	21.30899608
00017	ประ	1079	0.80782822	22.11682430
00018	นี้	1066	0.79809535	22.91491965
00019	กำ	868	0.64985625	23.56477590
00020	คน	856	0.64087206	24.20564796
00021	ผู้	844	0.63188787	24.83753533
00022	กัน	827	0.61916027	25.45669610
00023	แล้ว	827	0.61916027	26.07585637
00024	แต่	827	0.61916027	26.69501664
00025	จาก	788	0.58996166	27.28497830
00026	อย่าง	775	0.58022879	27.86520709
00027	นั้น	766	0.57349065	28.43869774
00028	อยู่	742	0.55552228	28.99422022
00029	กับ	691	0.51733948	29.51155950
00030	ต้อง	662	0.49562769	30.00718719
00031	ทาง	612	0.45819357	30.46538676
00032	หรือ	561	0.42001078	30.88539154
00033	งาม	549	0.41102659	31.29641813
00034	ด้วย	549	0.41102659	31.70744472
00035	ใจ	531	0.39755032	32.10499503
00036	ขึ้น	497	0.37209511	32.47709014
00037	ถึง	495	0.37059774	32.84768788
00038	ต่อ	481	0.36011619	33.20780407
00039	เข้า	475	0.35562410	33.56342817
00040	รับ	462	0.34589123	33.90931940
00041	น้ำ	460	0.34439386	34.25371326
00042	ยัง	460	0.34439386	34.59810712
00043	เรา	458	0.34289650	34.94100362
00044	มาก	456	0.34139913	35.28240275
00045	โดย	451	0.33765572	35.62005847
00046	ทั้ง	442	0.33091758	35.95097605
00047	หน้า	430	0.32193889	36.27290944
00048	วัน	420	0.31444657	36.58735601
00049	ซึ่ง	414	0.30995448	36.89731049
00050	ออก	412	0.30845711	37.20576760
00051	คุณ	410	0.30695975	37.51272735
00052	ใช้	408	0.30546238	37.81818973
00053	ดี	396	0.29647819	38.11466792
00054	เขา	390	0.29198610	38.40665402
00055	ตัว	383	0.28674532	38.69339984
00056	ตาม	381	0.28524796	38.9784790
00057	คือ	380	0.28449928	39.26314658
00058	รู้	380	0.28449928	39.54764586
00059	เพราะ	379	0.28375059	39.83139645
00060	จำ	372	0.27950982	40.10990627
00061	เมื่อ	368	0.27551507	40.38542136
00062	ปี	366	0.27401772	40.65943905
00063	เรื่อง	364	0.27252036	40.93195944
00064	อีก	362	0.27102299	41.20298243
00065	ไทย	359	0.26877695	41.47175938
00066	บ้าน	340	0.25455198	41.72631136
00067	พระ	327	0.24481911	41.97113047
00068	เห็น	325	0.24332175	42.21445222
00069	ทุก	313	0.23433756	42.44878918
00070	ผม	305	0.22834810	42.67715788
00071	จึง	303	0.22685074	42.90398862
00072	มัน	302	0.22610206	43.13009068
00073	เทศ	301	0.22535337	43.35544405
00074	กระ	299	0.22385601	43.57930006
00075	หนึ่ง	299	0.22385601	43.80315607
00076	เสีย	297	0.22235864	44.02551471
00077	รถ	296	0.22160996	44.24712467
00078	เงิน	294	0.22011260	44.46723727
00079	กว่า	292	0.21861523	44.68585250
00080	ตั้ง	290	0.21711787	44.902097037
00081	หา	288	0.21562050	45.11859087
00082	เพื่อ	287	0.21487182	45.33346269
00083	ท่าน	282	0.21112841	45.54459110
00084	ลง	280	0.20963104	45.75422214
00085	เมือง	277	0.20738500	45.96160714
00086	นัก	267	0.19989817	46.16150531
00087	นาย	259	0.19390872	46.35541403
00088	หลาย	257	0.19241135	46.54782538
00089	ไว้	255	0.19091399	46.73813937
00090	เอา	254	0.19016530	46.92890467
00091	เวลา	253	0.18941662	47.11832129
00092	เกิด	244	0.18267848	47.30099977
00093	ละ	242	0.18118112	47.48218089
00094	ผล	239	0.17893507	47.66111576
00095	จัด	238	0.17818639	47.83930235
00096	ดู	236	0.17668902	48.01599187
00097	ชน	230	0.17219693	48.18818830
00098	ถ้า	229	0.17144825	48.35963655

ลำดับที่	คำไทย	ความถี่	เปอร์เซ็นต์	เปอร์เซ็นต์สะสม
00099	จน	229	0.17144825	48.53108480
00100	ส่วน	229	0.17144825	48.70253305
00101	บาง	227	0.16995088	48.87248393
00102	ใหญ่	227	0.16995088	49.04243481
00103	เอง	221	0.16545879	49.20789860
00104	นำ	220	0.16471011	49.37260371
00105	เลย	216	0.16171538	49.53431909
00106	ครั้ง	214	0.16021801	49.69453710
00107	หนัง	213	0.15946933	49.85400643
00108	ก่อน	212	0.15872065	50.01272708
00109	เดิน	206	0.15422855	50.16695563
00110	เจ้า	206	0.15422855	50.31118418
10111	แห่ง	202	0.15123382	50.47241800
10112	ถูก	199	0.14898778	50.62140578
10113	เครื่อง	197	0.14249041	50.76889619
10114	ส่ง	196	0.14674178	50.91563792
10115	เดียว	196	0.14674173	51.06237965
10116	คำ	194	0.14524436	51.20762401
10117	คิด	194	0.14524436	51.35286837
10118	ลูก	194	0.14524436	51.49811273
10119	แบบ	194	0.14524436	51.64335709
10120	ราช	191	0.14299832	51.78635541
10121	เคย	189	0.14150095	51.92785636
10122	ค้า	187	0.14000359	52.06785995
10123	พอ	187	0.14000359	52.20786354
10124	ขอ	186	0.13925491	52.34711845
10125	ลัง	181	0.13551149	52.48262994
10126	สุด	181	0.13551149	52.61814143
10127	เช่น	181	0.13551149	52.75365292
00128	แม่	178	0.13326545	52.88691837
00129	กรรม	177	0.13251677	53.01943514
00130	จริง	177	0.13251677	53.15195191
00131	เรียน	177	0.13251677	53.28446868
00132	ข้อ	176	0.13176808	53.41623676
00133	สามารถ	176	0.13176808	53.54800484
00134	ชาติ	175	0.13101940	53.67902424
00135	ดัง	174	0.13027072	53.80929496
00136	ยาก	174	0.13027072	53.93956568
00137	อะไร	171	0.12802467	54.06759035
00138	เท่า	169	0.12652321	54.19411766
00139	รัก	168	0.12577862	54.31989623
00140	รัฐ	168	0.12577862	54.44567490
00141	ติด	167	0.12502994	54.57070484
00142	ประชา	166	0.12428126	54.69498610
00143	หมาย	165	0.12353258	54.81851868
00144	ต่ำ	163	0.12203527	54.94055389
00145	คง	162	0.12128653	55.06184042
00146	กล่าว	162	0.12128653	55.18321695
00147	ชื่อ	160	0.11978917	55.30291612
00148	ต้น	159	0.11904048	55.42195660
00149	สร้าง	159	0.11904048	55.54099708

ลำดับที่	คำไทย	ความถี่	เปอร์เซ็นต์	เปอร์เซ็นต์สะสม
00150	หลัง	158	0.11829180	55.65928888
00151	ราย	158	0.11829180	55.77758068
00152	แก่	158	.0.11829180	55.39587248
00153	สือ	157	0.117544312	56.01341560
00154	ต่าง	156	0.11679444	56.13021004
00155	ดิน	156	0.11679444	56.24700448
00156	ภาพ	155	0.11604575	56.36305023
00157	สอง	155	0.11604575	56.47909598
00158	มหา	154	0.11529707	56.59439305
00159	พวก	154	0.11529707	56.70969012
00160	ต่าง ๆ	152	0.11379971	56.82348983
00161	ตอน	152	0.11379971	56.93728954
00162	สำหรับ	152	0.11379971	57.05108925
00163	เหมือน	151	0.11305103	57.16414028
00164	พ่อ	150	0.11230234	57.27644262
00165	ควร	148	0.11080498	57.38724760
00166	บริษัท	148	0.11080498	57.49805258
00167	เสียง	148	0.11080498	57.60885756
00168	อัน	148	0.11080498	57.71966254
00169	น้อย	147	0.11005630	57.82971384
00170	ระ	147	0.11005630	57.93977514
00171	ฉัน	146	0.10930761	58.04908275
00172	ร่วม	146	0.10930761	58.15889036
00173	หมด	146	0.10930761	58.26769797
00174	อาจ	146	0.10930761	58.37700558
00175	ค่ำ	145	0.10855893	58.48556451
00176	ไร	145	0.10855893	58.59412344
00177	ปัญหา	144	0.10781025	58.70193369
00178	ที่	143	0.10706157	58.80899526
00179	กิน	143	0.10706157	58.91605683
00180	ด้าน	142	0.10631288	59.02236971
00181	น่า	142	0.10631288	59.12868259
00182	หัว	142	0.10631288	59.23499547
00183	เพียง	142	0.10631288	59.34130835
00184	ชาว	141	0.10556420	59.44687255
00185	พูด	141	0.10556420	59.55243675
00186	ธรรม	138	0.10331816	59.65575491
00187	นอก	137	0.10256947	59.75832438
00188	ใคร	137	0.10256947	59.86089385
00189	ข่าว	136	0.10182079	59.96271464
00190	ใหม่	135	0.10107211	60.06378675
00191	ครับ	133	0.09957474	60.16336149
00192	มือ	133	0.09957474	60.26293632
00193	โรง	133	0.09957474	60.36251097
00194	ขาย	130	0.09732870	60.45983967
00195	คณะ	129	0.09658007	60.55641968
00196	ตา	129	0.09658001	60.65299969
00197	ยิ่ง	129	0.09658001	60.74957970
00198	ช่วย	127	0.09508265	60.84466235
00199	ร้อย	127	0.09508265	60.93974500
00200	เรือ	127	0.09508265	67.03482765

ผลการวิจัยดังกล่าวสามารถนำตัวอักษรมาแจกแจงความถี่ของการใช้งาน โดยแยกแยะเฉพาะตัวอักษร (ดังตารางที่ 4) และพบว่าตัวอักษรที่ใช้งานกันมากที่สุดคือ า ใช้ถึง 7.1 เปอร์เซ็นต์ และ น ใช้สูงถึง 6.2 เปอร์เซ็นท์ สถิติของตัวอักษรและคำนี้สามารถนำไปใช้ในการสร้าง Spelling check dictionary ซึ่งเราคงเคยเห็นของเวิร์ดสตาร์มาแล้วว่า ทำไมจำนวนคำศัพท์ที่เก็บได้จำนวนมาก แต่เก็บในพื้นที่ดิสค์น้อย ทั้งนี้เพราะมีการใช้เทคนิคในการลดข้อมูลของคำศัพท์ซึ่งมีหลายเทคนิค ผู้เขียนขอให้พิจารณาจากการนำเอาสถิติของตัวอักษร มาสร้างรหัสตามทฤษฎี huffman คือสร้างรหัสเพื่อลดขนาดลง ผลที่ได้จะใช้รหัสแทนภาษาไทยดังตารางที่ 5

การลดข้อมูล(data compression) มีหลายเทคนิคโดยเฉพาะ หากต้องการพัฒนาการตรวจสอบคำในเวิร์ดโปรเซสซิ่ง จะต้องศึกษาเทคนิคต่างๆ เหล่านี้ โดยนำข้อมูลมาเปรียบเทียบ เพื่อให้การเปรียบเทียบคำศัพท์ได้รวดเร็วมาก และใช้พื้นที่ในดิสเกตต์น้อย การวิจัยประยุกต์ดังกล่าว จึงเป็นเรื่องที่จะทำให้การใช้ภาษาไทยมีประสิทธิภาพยิ่งขึ้น

นอกจากนี้ สถิติของคำไทยยังนำมาใช้ในการตัดแบ่งคำไทย (thai word (syllable) saparator) ทั้งนี้เพราะกฎเกณฑ์การตัดแบ่งแยกคำไทยไม่ตายตัว จำเป็นต้องใช้สถิติประกอบ โดยสร้างกฎเกณฑ์ต่างๆ เพื่อทดสอบ คณะผู้เขียนได้ทำการวิจัยโดยใช้ข้อมูลพื้นฐานประกอบกับหลักเกณฑ์ทางภาษาศาสตร์จำนวนประมาณ 20 กฎ สามารถใช้งานได้ด้วยเปอร์เซ็นต์สูงถึงกว่า 95 เปอร์เซ็นต์ อย่างไรก็ตาม สถิติพื้นฐานที่ได้จากการวิเคราะห์คำไทยในเบื้องแรกนี้คือ หาสถิติเกี่ยวกับลำดับของตัวอักษร: เช่น ตัวอักษรตัวแรกและตัวอักษรตัวสุดท้าย เช่น ผ ฝ เ แ โ ใ ไ เป็นตัวอักษรตัวแรกที่มีสถิติสูงมาก ดังตารางที่ 6 ส่วนตัวอักษรบางตัวควรจะอยู่ตัวสุดท้ายของคำเช่น ์ (การันต์) เป็นต้น โดยตารางแสดงสถิติตัวอักษรตัวท้าย ดังตารางที่ 7

ในการสร้างฐานความรอบรู้เกี่ยวกับการเข้าใจคำไทย หรือทำการตรวจสอบโครงสร้างคำไทย จึงได้สถิติรูปแบบคำไทย โดยแยกกลุ่มของตัวสระ พยัญชนะ วรรณยุกต์ โดยให้ C แทนพยัญชนะต้น S แทนตัวสะกด V แทนสระ T แทนวรรณยุกต์ G
การันต์ เช่น กิน แทนเป็น GVS, สิ้น
ด้วย CVTS, ศิลป์ แทนด้วย CVSSG,
แทนด้วย CCVS รูปแบบของคำไทย
แบบต่างๆ แสดงได้ดังตารางที่ 8

ตารางที่ 4 แสดงการแจกแจงความถี่ตัวอักษร

ตัวอักษร	จำนวนตัว	เปอร์เซ็นต์
	670	0.1446
ก	19963	4.3092
ข	5599	1.2086
ค	8332	1.7985
ฆ	111	0.0240
ง	19176	4.1393
จ	7259	1.5669
ฉ	519	0.1120
ช	4366	0.9424
ซ	1049	0.2264
ฌ	1	0.0002
ญ	1239	0.2674
ฎ	98	0.0212
ฏ	210	0.0453
ฐ	465	0.1004
ฑ	50	0.0108
ฒ	167	0.0360
ณ	1663	0.3590
ด	10813	2.3341
ต	8407	1.8147
ถ	2386	0.5150
ท	11004	2.3753
ธ	1361	0.2938
น	28917	6.2420
บ	7892	1.7036
ป	8266	1.7843
ผ	2272	0.4904
ฝ	317	0.0684
พ	5548	1.1976
ฟ	505	0.1090
ภ	1101	0.2377
ม	16406	3.5414
ย	13854	2.9905
ร	22720	4.9043

ตัวอักษร	จำนวนตัว	เปอร์เซ็น
ฤ	162	0.08
ล	10365	2.23
ว	12882	2.78
ศ	1615	0.34
ษ	1139	0.24
ส	8511	1.83
ห	9802	2.11
ฬ	63	0.01
อ	20346	4.39
ฮ	120	0.02
ะ	8765	1.89
า	32908	7.10
ำ	3658	0.78
เ	18084	3.9
แ	6846	0.4
โ	2321	0.5
ใ	5346	1.1
ไ	7068	1.5
ๆ	825	0.1
ฯ	89	0.0
ุ	4180	0.9
ู	3993	0.8
ิ	8887	1.9
ี	12388	2.6
ึ	2650	0.5
ื	5443	1.1
ั	13129	2.8
็	4848	1.0
่	22600	4.8
้	16684	8.6
๊	165	0.0
๋	159	0.0
์	2567	0.5
ะ	1953	0.4

ตารางที่ 5 การใช้สถิติสร้างรหัสแบบ huffman ทำให้ข้อมูลลดลงเหลือเพียง 5.27 บิท

161	ก	431	11110	5	195	ล	224	111111	6
162	ข	121	010100	6	196	ว	278	01100	5
163	ค	189	110000	6	197	ศ	35	01010101	8
164	ฆ	2	1110001110111	13	198	ษ	24	111000110	9
165	ง	414	11101	5	199	ส	184	110010	6
166	จ	157	100100	6	200	ห	211	111001	6
167	ฉ	11	010101111	10	201	ฬ	1	0101011100111	13
168	ช	94	1100111	7	202	อ	439	0000	4
169	ซ	23	111000000	9	203	ฮ	2	010101110001	12
170	ฌ	1	11100011101101	14	204	ะ	109	110100	6
171	ญ	27	00010000	8	205				
172	ฎ	2	010101110000	12	206	า	710	1010	4
173	ฏ	4	01010100010	11	207	ำ	79	1001010	7
174	ฐ	10	0101011101	10	208	เ	390	11011	5
175	ฑ	1	0101011100110	13	209	แ	148	011110	6
176	ฒ	4	01010100011	11	210	โ	50	11100010	8
177	ณ	36	01010110	8	211	ใ	115	000101	6
178	ด	233	00011	5	212	ไ	153	011111	6
179	ต	181	110001	6	213	ฯ	18	010101001	9
180	ถ	51	11111010	8	214	ๆ	2	010101110010	12
181	ท	237	00100	5	215	ุ	90	1100110	7
182	ธ	29	00010001	8	216	ู	86	1991911	7
183	น	624	1000	4	217	ั	192	119191	6
184	บ	170	100110	6	218	ิ	267	01011	5
185	ป	178	100111	6	219	ี	57	0001001	7
186	ผ	49	11100001	8	220	ึ	117	001010	6
187	ฝ	7	0101010000	10	221	ื	283	01101	5
188	พ	120	001011	6	222	•			
189	ฟ	11	0101011111	10	223	็	105	1111100	7
190	ภ	24	111000001	9	224	่	489	0011	4
191	ม	354	10110	5	225	้	360	10111	5
192	ย	299	01110	5	226	๊	3	111000111001	12
193	ร	490	0100	4	227	๋	3	111000111010	12
194	ฤ	3	111000111000	12	228	์	55	11111011	8

ตารางที่ 6 แสดงการแจกแจงตัวอักษรที่ใช้เป็นตัวแรก

ตัวอักษร	จำนวนตัวที่ใช้ทั้งหมด	จำนวนตัวแรก	เปอร์เซ็นต์
	670	0	0.0000
ก	19963	9648	48.3294
ข	5599	3955	70.6376
ค	8332	5804	69.6591
ฆ	111	16	14.4144
ง	19176	733	3.8225
จ	7259	4989	68.7285
ฉ	519	346	66.6667
ช	4366	2269	51.9698
ซ	1049	793	75.5958
ฌ	1	0	0.0000
ญ	1239	57	4.6005
ฎ	98	3	3.0612
ฏ	210	0	0.0000
ฐ	465	91	19.5699
ฑ	50	0	0.0000
ฒ	167	0	0.0000
ณ	1663	57	3.4275
ด	10813	2494	23.0648
ต	8407	4515	53.7052
ถ	2386	1512	63.3697
ท	11004	7918	71.9557
ธ	1361	478	27.7737
น	28917	5155	17.8269
บ	7892	2441	30.9301
ป	8266	3414	41.3017
ผ	2272	2007	88.3363
ฝ	317	306	96.5300
พ	5548	2984	53.7851
ฟ	505	260	51.4851
ภ	1101	649	58.9464
ม	16406	5038	30.7083
ย	13854	1697	12.2492
ร	22720	4359	19.1857
ฤ	162	56	34.5679
ล	10365	1924	18.5625
ว	12882	3127	24.2742
ศ	1615	657	40.6811
ษ	1139	0	0.0000
ส	8511	6069	71.3077
ห	9802	4846	49.4389
ฬ	63	0	0.0000
อ	20346	5347	26.2803

ตัวอักษร	จำนวนตัวที่ใช้ทั้งหมด	จำนวนตัวสุดท้าย	เปอร์เซ็นต์
ฮ	120	50	41.6667
ะ	8765	0	1.0000
า	32908	0	1.0000
ำ	3658	0	1.0000
เ	18084	16915	93.5357
แ	6846	6690	97.7213
โ	2321	1875	80.7841
ใ	5346	5343	99.9439
ไ	7068	6766	95.7272
ๆ	825	0	0.0000
ฯ	89	15	16.8539
ุ	4180	0	0.0000
็	3992	0	0.0000
ิ	8987	0	0.0000
ี	12388	0	0.0000
ึ	2650	0	0.0000
ื	5443	0	0.0000
ั	13129	0	0.0000
ู	4848	0	0.0000
่	22600	0	0.0000
้	16684	0	0.0000
๊	165	0	0.0000
๋	159	0	0.0000
์	2567	0	0.0000
ะ	1953	0	0.0000

ตัวอักษร	จำนวนตัวที่ใช้ทั้งหมด	จำนวนตัวสุดท้าย	เปอร์เซ็นต์
	670	441	65.8209
ก	19963	7095	35.5408
ข	5599	189	3.3756
ค	8332	409	4.9088
ฆ	111	25	22.5225
ง	19176	17099	89.1688
จ	7259	1233	16.9850
ฉ	519	2	0.3854
ช	4366	325	7.4439
ซ	1049	15	1.4299
ฌ	1	0	0.0000
ญ	1239	389	31.3156
ฎ	98	54	55.1020
ฏ	210	60	28.5714
ฐ	465	264	56.7742
ฑ	50	0	0.0000
ฒ	167	2	1.1976
ณ	1663	750	45.0992

ตัวอักษร	จำนวนตัวที่ใช้ทั้งหมด	จำนวนตัวสุดท้าย	เปอร์เซ็นต์
ด	10813	4242	39.2768
ต	8407	642	7.6365
ถ	2386	502	21.0394
ท	11004	567	5.1527
ธ	1361	139	10.2131
น	28917	18347	61.4915
บ	7892	4225	53.5352
ป	8266	1661	20.0944
ผ	2272	1	0.0440
ฝ	317	1	0.3155
พ	5548	503	9.0663
ฟ	505	146	28.9109
ภ	1101	8	0.7266
ม	16406	5883	35.8588
ย	13854	6348	45.8207
ร	22720	4857	21.3776
ฤ	162	0	0.0000
ล	10365	1110	10.7091
ว	12882	3371	26.1683
ศ	1615	506	31.3313
ษ	1139	202	17.7349
ส	8511	273	3.2076
ห	9802	8	0.0816
ฬ	63	0	0.0000
อ	20346	4745	23.3215
ฮ	120	0	0.0000
ะ	8765	7386	84.2670
า	32908	10416	31.6519
ำ	3658	2217	60.6069
เ	18084	0	0.0000
แ	6846	0	0.0000
โ	2321	0	0.0000
ใ	5346	0	0.0000
ไ	7068	0	0.0000
ๆ	825	825	100.0000
ฯ	89	74	83.1461
ุ	4180	374	8.9474
็	3993	475	11.8958
ิ	8987	952	10.7123
ี	12388	3708	29.9322
ึ	2650	17	0.6415
ื	5443	0	0.0000
ั	13129	1	0.0076
ู	4848	1552	32.0132
่	22600	8491	37.5708
้	16684	7713	46.2299
๊	165	6	3.6364
๋	159	8	5.0314
์	2567	2397	93.3775
ะ	1953	0	0.0000

ตารางที่ 8 รูปแบบคำไทยและความถี่ที่ใช้

CCCG	159	CVSV	1166
CCCS	450	CVSUS	739
CCCSV	114	CVSVSS	229
CCCV	123	CVSVSSG	187
CCCVS	137	CVSUSSV	170
CCS	991	CVT	6934
CCTV	846	CVTS	4457
CCTVS	778	CVTSG	105
CCV	2727	CVTV	353
CCVCV	168	CCVC	227
CCVS	6241	CVVCS	153
CCVSS	233	VC	4739
CCVSV	101	VCC	4194
CCVSVS	512	VCCG	127
CCVSVSV	109	VCCS	598
CCVTS	813	VCCSG	121
CGCG	104	VCCSS	108
CS	4652	VCCT	519
CSTVS	791	VCCTS	111
CSVS	151	VCCTV	128
CSVT	742	VCCV	434
CTS	542	VCCVS	479
CTV	3644	VCCUSS	374
CTVS	6127	VCCVTSS	106
CTVSG	211	VCCVTVS	252
CV	12272	VCCVV	518
CVCS	150	VCS	380
CVCSVS	203	VCT	7859
CVCV	118	VCTS	2033
CVCVS	536	VCTV	1139
CVS	29497	VCV	2806
CVSS	437	VCVS	4805
CVSSG	503	VCVSS	1838
CVSSS	200	VCVTS	266
CVSSSG	150	VCVTSS	392
CVSSSV	110	VCVTV	864
CVSSV	896	VCVTVS	646
CVSSVS	241	VCCV	153
CVSSVSVS	168	VVSV	261

คำไทยที่ใช้นี้ มีขนาดของคำไม่แน่นอน
รเขียนโปรแกรมตัดแบ่งแยกคำไทยวิธีหนึ่ง
ใช้หลักการที่เรียกว่า กรอบช่องหน้าต่าง
มกฎเกณฑ์การตัดแบ่ง โดยทำเสมือนเป็น
อบช่องหน้าต่างไปตามประโยค แล้วใช้
เกณฑ์วินิจฉัย เพื่อเพิ่มความรวดเร็วของ
รตัดแบ่ง ซึ่งเป็นวิธีของซอฟท์แวร์ ซึ่งจะยังไม่กล่าวในที่นี้ แต่สิ่งที่จะมีประโยชน์อย่างหนึ่งคือ สถิติการแจกแจงขนาดของคำ ซึ่งแสดงให้เห็นดังตารางที่ 9

การเสนอสถิติพื้นฐานนี้ เป็นเพียงข้อมูลที่จะนำมาใช้ในการสร้างระบบซอฟท์แวร์ต่างๆ ในตอนต่อไปจะเสนอ หลักการทางภาษาศาสตร์กับการใช้ไมโครคอมพิวเตอร์

ตารางที่ 9 แสดงการแจกแจงความถี่ของขนาดของคำ

ขนาดของคำ	จำนวนคำ	เปอร์เซ็นต์
1	55	0.0412
2	21796	16.3183
3	61391	45.2628
4	31012	23.2181
5	11532	8.6338
6	4191	3.1677
7	1857	1.3903
8	954	0.7142
9	287	0.2149
10	244	0.1827
11	152	0.1138
12	36	0.0270
13	30	0.0225
14	20	0.0150
15	0.	0.0000

กิตติกรรมประกาศ

ผู้เขียนขอขอบคุณบริษัทถาวรคอมพิวเตอร์ โดยเฉพาะฝ่ายวิจัยและพัฒนา ที่ได้เสนอความคิดเห็นตอบสนองแนวความคิดของผู้เขียนหลายเรื่อง โดยเฉพาะในเรื่องการจัดการข้อมูลระบบไทย 25 บรรทัด เพื่อให้วิชาการด้านนี้มีการพัฒนา และทดสอบแนวความคิดถึงผลดีผลเสียของการจัดการข้อมูลแบบต่างๆ

หนังสืออ้างอิง

1. ยืน ภู่วรวรรณ และคณะ "รายงานการวิจัยเรื่อง การประมวลผลข้อมูลภาษาไทยด้วยไมโครคอมพิวเตอร์" รายงานเสนอคณะกรรมการสภาวิจัยแห่งชาติ เพื่อรับรางวัลผลงานวิจัยดีเด่นที่เป็นประโยชน์ต่อประเทศชาติ ประจำปี 2526

2. ยืน ภู่วรวรรณ และคณะ "อักกอริทึมคำไทยและการสังเคราะห์เสียงพูดภาษาไทย" รายงานเสนอคณะกรรมการสภาวิจัยแห่งชาติ เพื่อรับรางวัลผลงานวิจัยดีเด่นที่เป็นประโยชน์ต่อประเทศชาติ ประจำปี 2527